Random Portfolios for Evaluating Trading Strategies

Patrick Burns*

13th January 2006

Abstract

Random portfolios can provide a statistical test that a trading strategy performs better than chance. Each run of the strategy is compared to a number of matching random runs that are known to have zero skill. Importantly, this type of backtest shows periods of time when the strategy works and when it doesn't. Live portfolios can be monitored in this way as well. This allows informed decisions—such as changes in leverage—to be made in real-time.

1 Introduction

Random portfolios—portfolios that obey given constraints but ignore utility—are a powerful tool in finance. [Burns, 2004] discusses the use of random portfolios for measuring the performance of funds. The focus here is on using them to find a good trading strategy—a related but distinct task.

A strategy has two parts: a means of predicting returns (alpha model), and a method of trading to try to take advantage of the alpha.

Performing a statistical test of the predictions is relatively easy. There are data snooping problems, but even so it is generally possible to have a good sense of whether or not a prediction method is picking up a signal. Random portfolios provide a rigorous test of the trading strategy as a whole—something that seems virtually impossible without random portfolios.

Suppose that we have the results of a trading strategy over some period of time. If we had a list of all of the possible trading paths that we might have taken, then we would know precisely how good our strategy was for the period, and for any sub-period. We would know that our strategy outperformed $x\%$ of the population of paths.

In our example the universe is of size 186 and the portfolios are of size 50. There are 6.89e45 ways of selecting the 50 stocks in the portfolio at the end of the trading period. For each of those sets of stocks there will be numerous

*This report can be found in the working papers section of the Burns Statistics website http://www.burns-stat.com/.

ways of selecting the number of shares for each stock that obey the constraints. For any given final portfolio there will be numerous paths to get there from the initial portfolio. The number of possible paths is finite, but such a large number that it is practically infinite.

But we don't need to have all of the paths in order to evaluate our strategy. If we generate a random subset of the paths, then we can make statistical statements about the quality of the strategy. All introductory statistics books discuss sampling from a population, and that is merely what we are doing. A few thousand paths is the most we would ever need for practical purposes.

R [R Development Core Team, 2005] and the POP Portfolio Construction Suite [Burns Statistics, 2005] were used for computations in this paper.

2 Example Data

A specific example is used to illustrate the use of random portfolios for evaluating a strategy.

The universe of stocks is 186 US equities that are an unsystematic mixture of large caps and small caps. Daily data are used beginning at the start of 1996. The first 500 days are used to estimate the variance matrix for the first portfolio optimization, and to evaluate the strength of the prediction. The next 1000 days are the period in which trading occurs. Data subsequent to these 1500 days is left untouched so it could be used for testing a final strategy.

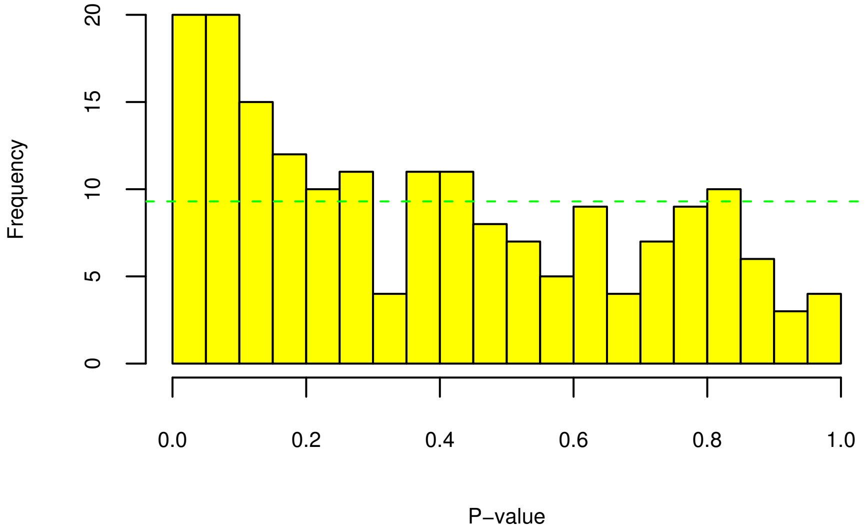
The alpha model for each stock is the equally weighted mean of the returns on the previous 26 trading days minus the equally weighted mean of the returns on the previous 12 days. That is, it is a Moving Average Convergence Divergence estimate. Aficionados of MACD usually use exponential weights though.

3 Evaluating Prediction

The first step in evaluating a strategy is to test the prediction process. One common approach is to do a sign test—a "success" is scored if the prediction and the realized return are either both above their median or both below their median. It's a "failure" if one is below its median and the other is above its median. The binomial distribution is used to evaluate the probability.

We can also test if the Spearman correlation between the prediction and the realized returns is positive. The Spearman correlation uses the ranks rather than the actual data values. It is a slightly robust version of the usual (Pearson) correlation. The Spearman correlation doesn't react as strongly to outliers, but it is still affected a fair amount by outliers—as it should be in this setting. Besides exhibiting about the right amount of robustness, another reason to prefer the Spearman correlation is that the p-values from the test will be close to right even when the distributions are not very close to the normal distribution. Returns should not be assumed to follow the normal distribution.

Figure 1: P-values (one for each stock) of the Spearman correlation test for 2-day returns during the pre-trading and trading periods.



Figures 1 and 2 show the distribution of p-values from the Spearman test for predicting 2 and 5 days ahead over the first 1500 days of data, which includes the trading period. In both cases there is a disproportionate number of stocks that have small p-values. We are pleased.

As Figure 3 shows, the sign test need not necessarily agree with the corresponding Spearman test. The p-values can be substantially different, especially when the evidence is ambiguous.

A better approach to testing the prediction is to use the data before the trading period. This preserves the trading period from data snooping bias caused by searching for an adequate predictor. Figure 4 shows one test for this period. There is actually a deficit of stocks with small p-values. Clearly this predictor is not useful for all time periods. In a real case we would not have proceeded to testing the trading after seeing this—we would hone the predictor before moving on. (We would also have needed to use more data for the prediction testing period.)

4 Performance Evaluation

In the absence of random portfolios, it is hard to get a good sense of the quality of the strategy. The average return over the period is one obvious measure. However, this is unlikely to be a valid indicator of future returns—in fact, our example will show that it can be quite misleading.

Figure 2: P-values (one for each stock) of the Spearman correlation test for 5-day returns during the pre-trading and trading periods.

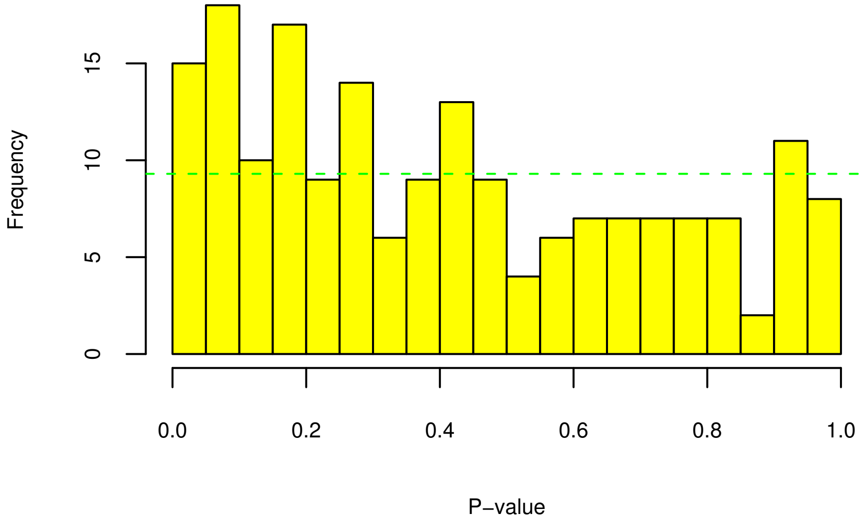


Figure 3: P-values of the Spearman test versus the sign test for 5-day returns during the pre-trading and trading periods.

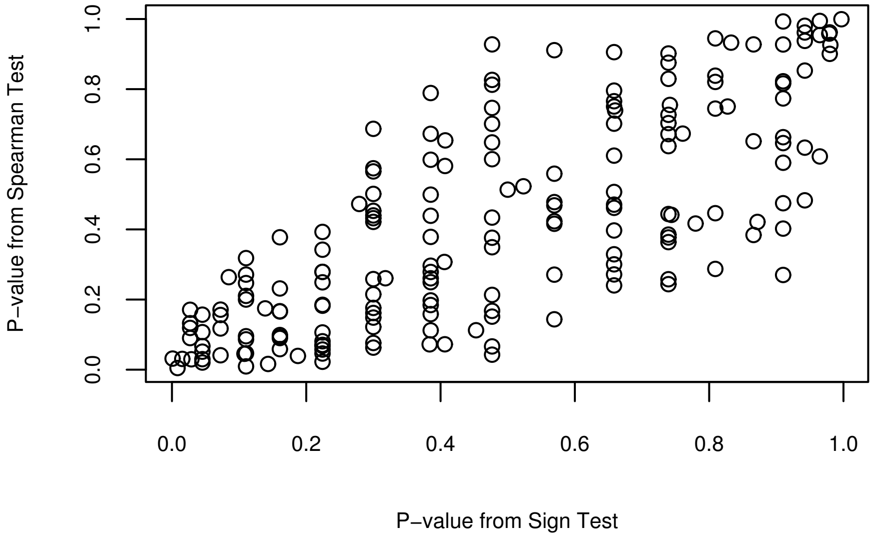
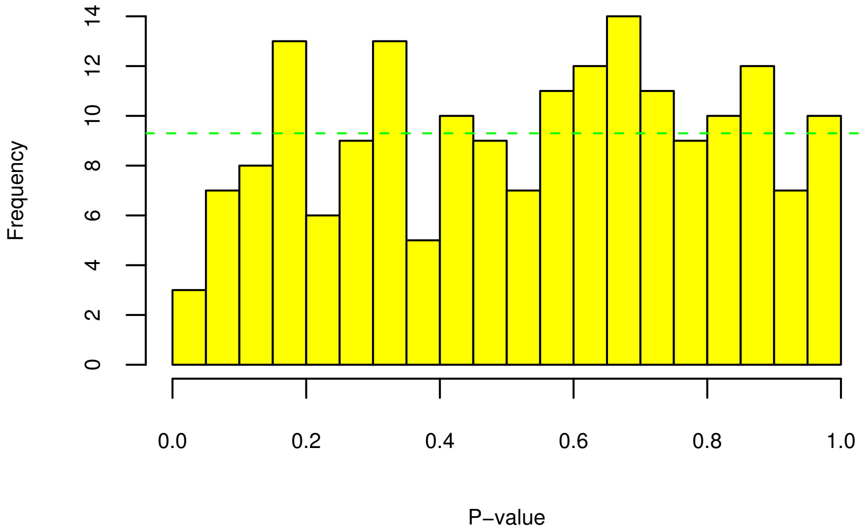


Figure 4: P-values (one for each stock) of the Spearman test for 5-day returns during the period before trading.



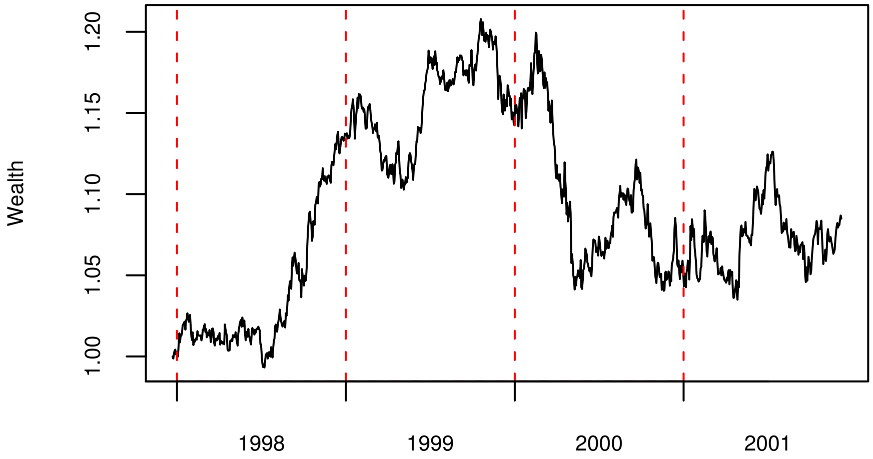
An optimization is performed each day in our test strategy—trades are done at the close of the day after the most recent data used in estimates. The objective is to maximize the information ratio with the constraint that turnover (buys plus sells) is limited to about 400% per year. The number of stocks in the portfolio is constrained to be between 45 and 50—in actuality, the portfolio is almost always of size 50. The portfolio is long-short with the net value held close to zero. It tries to have the absolute value of the net less than 5% of the gross value, and tries very hard to have it less than 10% of the gross. It attempts to keep the maximum weight of each asset in the portfolio less than 10% (where weight is the absolute size of a position divided by the gross of the portfolio). The variance matrix on each day is a statistical factor model built with the data from the previous 500 days.

4.1 Single Starting Portfolio

Figure 5 shows the wealth curve of the trading strategy that starts from a specific portfolio. The initial portfolio is roughly equally weighted in the (alphabetically) first 50 stocks in the universe with every second stock having a short position. The starting portfolio is an arbitrary portfolio that satisfies the constraints. The curve does not account for trading costs, but given that the turnover is constrained to be close to 400% per year, the effect of trading costs is easily assessed.

We want to generate random portfolios that mimic the actual optimization

Figure 5: The wealth generated from the trading strategy.



backtest. In essence we create a number of hypothetical fund managers that perform the same task that we do, but have no skill. If we outperform most of these hypothetical managers, that is evidence that we have skill. In fact, we can get an estimate of just how much skill we are exhibiting.

Here is an outline of how to create the portfolios for the hypothetical managers. The first step is to create a list, call it `exist_list`, with length equal to the number of random portfolios to be generated (100 in the example). Initialize each component of the list with the starting portfolio. Now loop over the trading times. The first thing is to update the expected returns and variance matrix. For each random path: get the existing portfolio; generate a random trade away from the existing portfolio; save the desired information about the new portfolio; put the new portfolio into the appropriate location of `exist_list`. (End loop over random paths, end loop over trading times.)

Figure 6 shows the wealth curves of 100 random portfolios with the same constraints as the optimization process. An easier way of seeing the pattern of the random portfolios is to plot a few quantiles at each point in time. The lines that are plotted are not individual portfolios, but switch portfolios from time to time. Figure 7 shows the actual optimization wealth relative to quantiles of the random portfolios. It has mediocre performance until mid-1998, at which point it clearly outperforms the random portfolios. It has poor performance at the dot-com collapse, and then recovers slightly.

It is strange that a strategy that gains only 8.5% in 4 years tests so well. Notice that the quantiles of the random portfolios are, in general, downward sloping. The original portfolio loses money over this time period, and its in-

Figure 6: Paths of 100 random portfolios with the same constraints as the optimization.

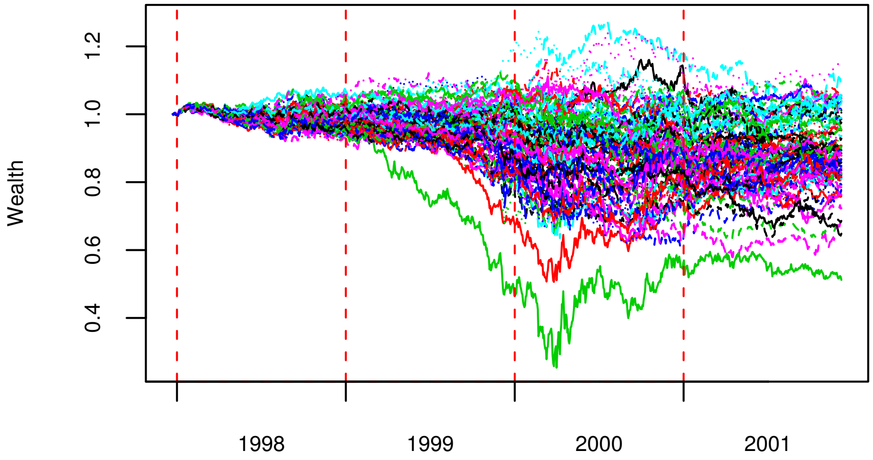


Figure 7: Random portfolio quantiles (minimum, 5%, 10%, 25%, 50%, 75%, 90%, 95%, maximum) in blue, and the actual optimization (in black).

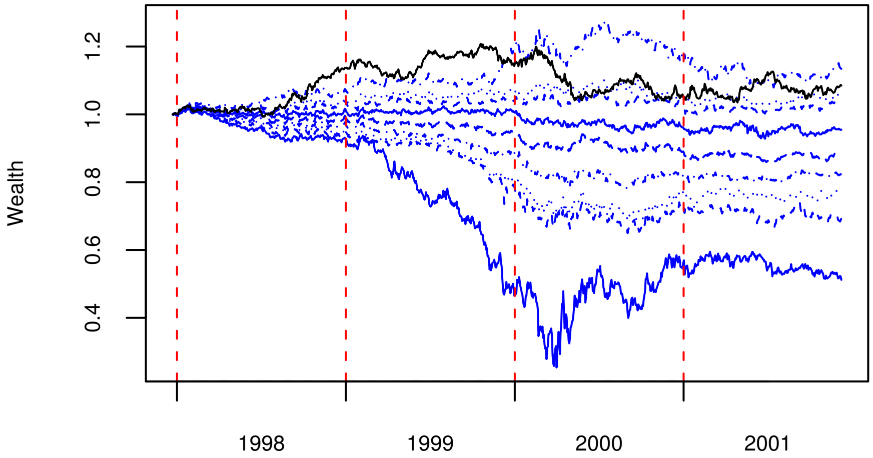
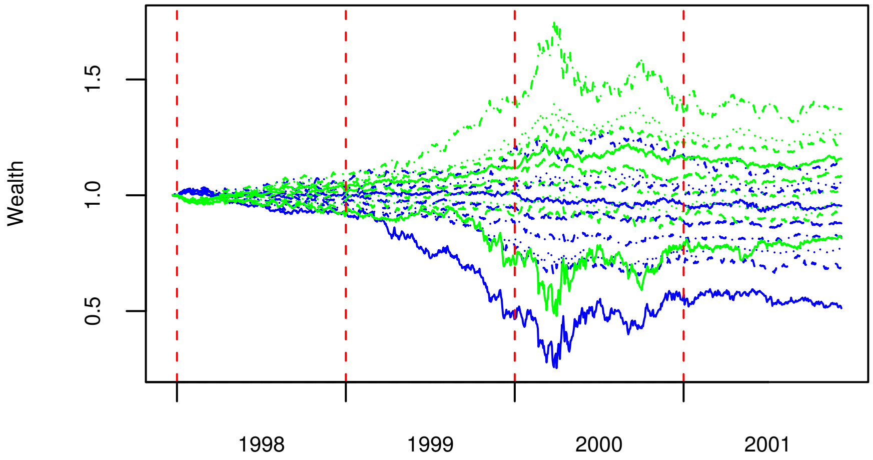


Figure 8: Quantiles of random portfolios starting with the specified portfolio (blue) and quantiles of random portfolios starting with the opposite of the specified portfolio (green).



fluence persists even though the portfolio is traded 1600% by the end of the period. This starting portfolio is a severe handicap.

We can see this by generating random portfolios that have the opposite starting portfolio—that is, the long positions are made short and the short positions made long. A comparison of the quantiles of the two sets of random portfolios is in Figure 8. Probably few would have guessed that the influence of the starting portfolio would persist so long. If the wealth curve of our strategy had to be compared with the random portfolios using the opposite starting portfolio, we certainly would not have found it to be exceptional.

Random portfolios that are long-short will be close to symmetric around no gain if changing the sign of each position of a random portfolio also satisfies the constraints. In the present case having a specific starting portfolio breaks that symmetry. The distributions represented in Figure 8 are close to mirror images of each other.

There are numerous ways in which the symmetry can be broken. For instance, if the range for the net value is not symmetric around zero.

4.2 P-value Control Charts

Perhaps more important than an evaluation of the strategy for the entire trading period is to identify sub-periods in which the strategy worked particularly well or poorly. On any trading day, the number of random portfolio returns that are larger than the return from the real portfolio is the key ingredient for the p-value of a test. The test is that the return from the portfolio is no larger than the mean random portfolio return. The p-values from individual days can be combined via Stouffer's method (see [Burns, 2004]) to get a smoother picture of when the strategy performed well.

Figure 9 shows 10-day non-overlapping p-values. There are points in time where the portfolio suddenly switches to a worse or a better state relative to the random portfolios. We would like the p-values to be no greater than 0.5, but in this case there are periods in which they approach 1 for some time. This means that the strategy is subject to significant drawdowns, and hence not particularly appetizing.

Another trait in Figure 9 is that the strategy seems to perform worse as time progresses. There are at least two possible explanations of this. One is that the alpha model loses power throughout the period—this could be either temporary or permanent. Another possibility is that the random portfolios are somehow systematically diverging from the actual strategy.

The volatility of the random portfolios in the final two years is in general significantly higher than the volatility of the optimized portfolio during that time. While the optimization does not formally have a constraint on volatility, the optimization process favors lower volatility. Constraining the random portfolios to have a volatility not much larger than the volatility of the optimized portfolio would probably provide a fairer assessment. Without such a constraint we would expect the p-values to drift higher over time.

While it is natural to suppose that the number of closing positions would be greater in the optimization than in the random portfolios, in actuality there were more closing positions in the random portfolios. So the random portfolios may be diverging faster from the original portfolio (which performs poorly) than the optimized strategy. A control chart generated from random portfolios that always have the same number of closing positions as the real portfolio shows less of a trend in the p-values. This eliminates the suspicion of the alpha model becoming worse over time.

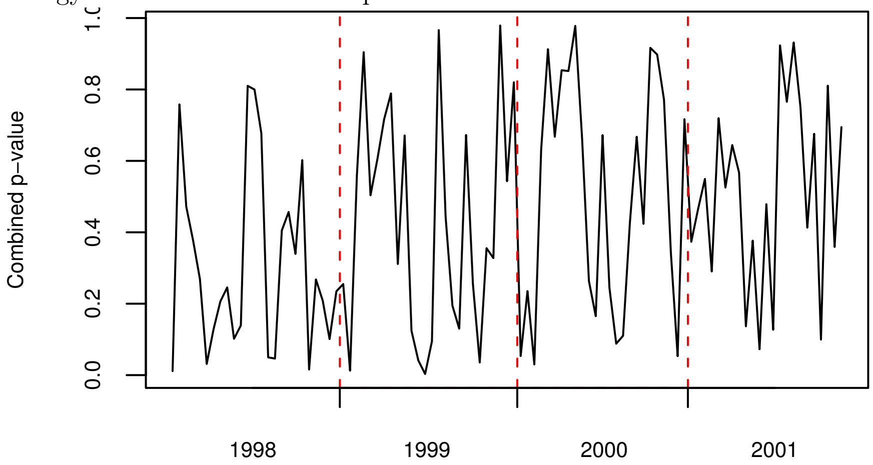
Plots like Figure 9 can be used in real-time to monitor if a portfolio is losing its edge. The plot could be used as a control chart to signal when the portfolio has gone "out of control" (either to the bad side or the good side). If the series of non-overlapping p-values is predictable, then it could be used to make decisions—for example, on changing the leverage.

4.3 Complete Evaluation Process

The wealth curve for the optimized portfolio of Section 4.1 has a p-value of 3% for the whole time period relative to the random paths that were generated. This is the p-value *given that we start with the portfolio that we did*. We aren't actually interested in that particular starting portfolio, we want to know how well the strategy performs starting from anywhere.

Here we outline our recommendation of the entire process.

Figure 9: P-values combined over 10 non-overlapping days for the returns of the strategy relative to the random portfolio returns.



- For each of several random starting portfolios perform the procedure given in Section 4.1.

- Examine the set of p-values that are obtained—one for each starting portfolio. Very few p-values should be larger than one-half, and you should be quite concerned if any are close to 1.

- Look at the p-value control chart for each run. Plotting multiple p-value paths in one chart could highlight times when the strategy does especially poorly (or well).

If the p-values are uniformly small—both for the whole trading period and within the trading period—then the strategy will be good.

Let's apply this to our example strategy. Figure 10 displays the wealth curves of the strategy for 20 randomly selected starting portfolios. Figure 11 shows the average wealth curve from these 20 runs. Since we know in this case that no skill is equivalent to zero gain, we can make some general observations. (Otherwise we could have plotted the wealth curve of the average random portfolios as well.) 1998 is a good year for the strategy, 1999 is about flat, and the first part of 2000 is bad. MACD is basically a momentum strategy. It makes sense that 1998 should be good, and that early 2000 (when the stock market was mean-reverting) should be bad. It is a bit puzzling that 1999 was not also a good year for the strategy.

Figure 10: Wealth curves of the example strategy from 20 random starting portfolios.

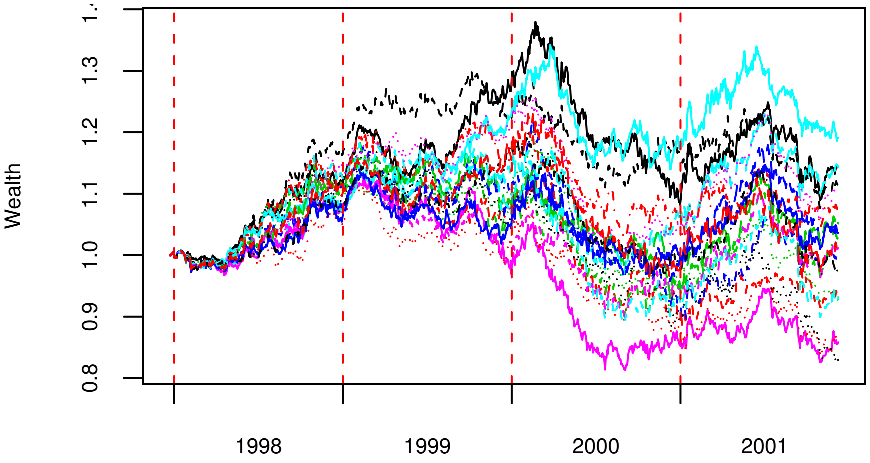


Figure 11: Wealth curve of the strategy averaged over the 20 starting portfolios.

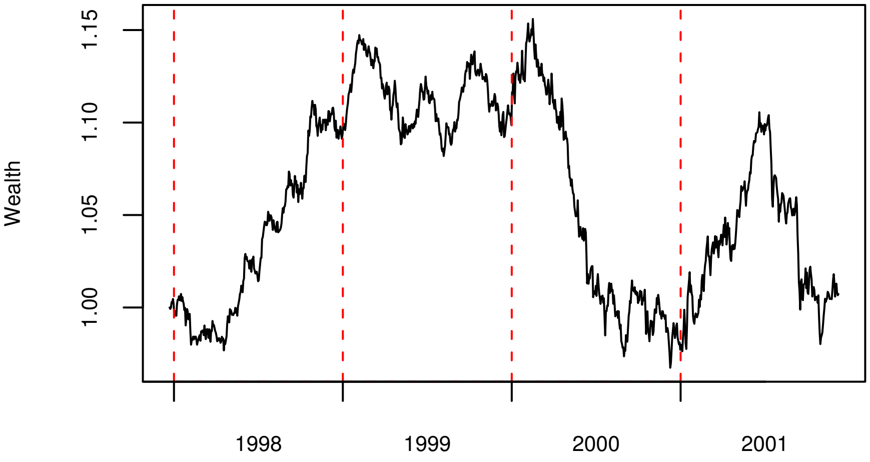
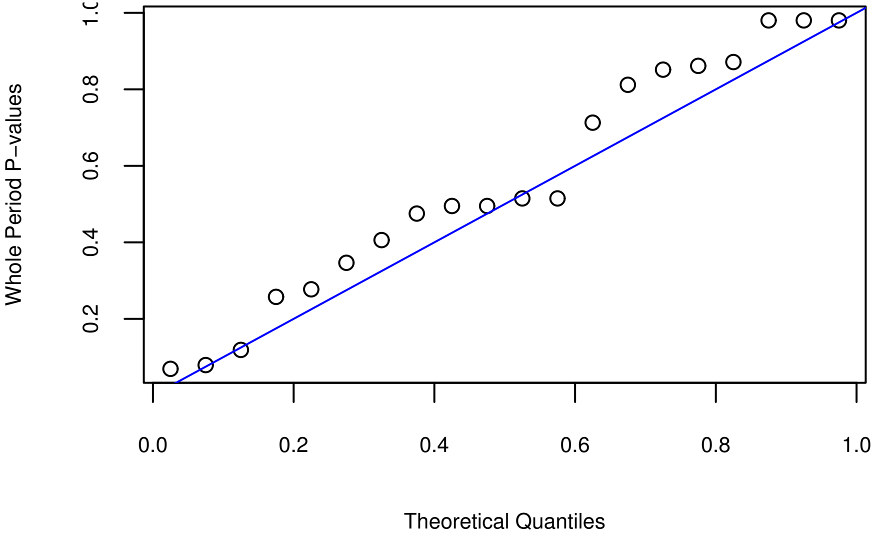


Figure 12: The 20 whole period p-values compared to the theoretical uniform distribution.



12 of the 20 wealth curves end the period with gains. But p-values are a more telling statistic. Figure 12 displays the 20 p-values for the whole trading period versus the expected values from the uniform distribution. We want the points to be below the line—this shows the strategy being slightly worse than no skill over this time period. (Constraining volatility in the random portfolios might have improved the results slightly.) This is a key plot, it is evidence that our strategy definitely should not be used. From Figure 10 it is clear that the strategy would have looked very good if the trading period were only 1998. Even though we have evidence that we have prediction power, we aren't using that to good advantage.

There are (at least) two ways to get a p-value from an optimization run and its associated random portfolios. The first is to count the number of random paths that outperform the optimized path—this is what is plotted in Figure 12. The second is to combine the daily p-values over the trading period. These are subtly different in meaning—is there outperformance over the period versus is there ever outperformance. Figure 13 compares these for the 20 random starting portfolios. The combined p-values are substantially smaller in this case. That means that the strategy has more days of being really good than really bad.

We also examine the p-values throughout the trading period for the 20 runs. Figure 14 plots the first and third quartiles of 10-day p-values (combined from daily p-values). Though noisy, there are clearly good and bad periods.

Our example strategy has an inefficiency. The same fraction of the value of the portfolio is traded each day. However, the value of trading is highly unlikely

Figure 13: Comparison of the whole period p-values to the combined daily p-values.

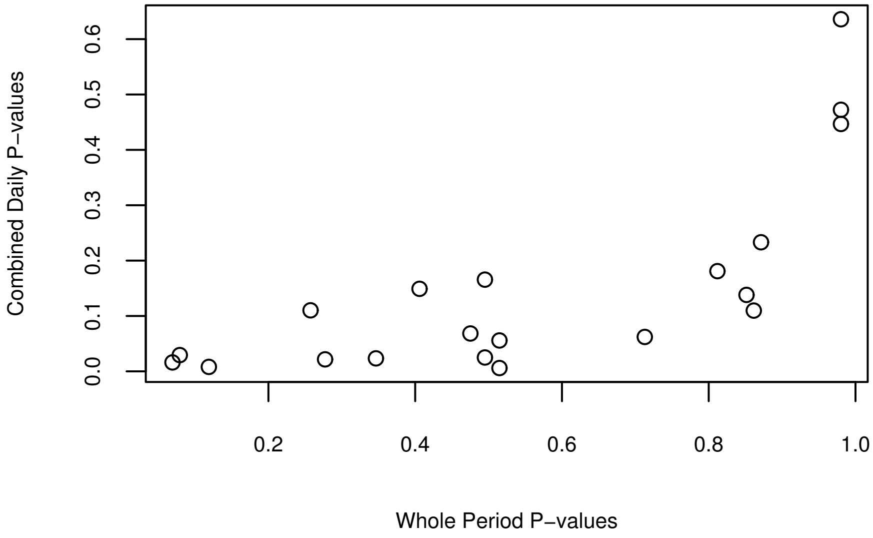
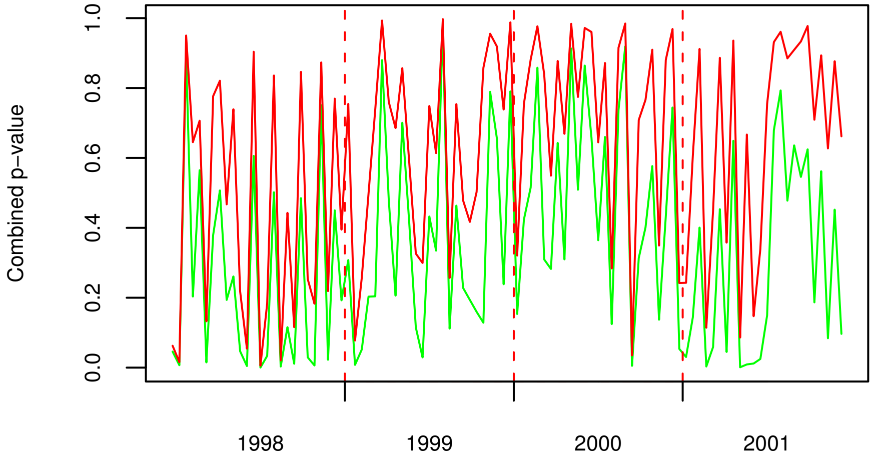


Figure 14: The first and third quartiles of the 10-day non-overlapping p-values from the 20 runs.



to be constant. A better approach is to use trading costs to limit the amount that is traded. More will be traded when the existing portfolio is expected to do poorly than when it is expected to do well. While getting the trading costs to be approximately right is non-trivial, it can be quite a valuable effort.

5 Comparing Competing Strategies

In the previous section we took rather a cynical view and asked if the strategy exhibited any value at all. Once you are in the position to believe that you have a strategy that does have value, you may want to compare it with another strategy to determine if either is significantly better.

If the constraints for the two strategies are the same, then a reasonable approach is to test the difference in returns from them. If the trading were daily, then a starting portfolio would be fixed, the two strategies would be run, and the data used in the test would be the differences of the daily returns. A t-test would be approximately correct, however the differences in returns would probably have longer tails than the normal distribution. A sign test or a signed-rank test may be more appropriate. P-values could be combined from tests based on different starting portfolios.

Random portfolios could be used in this case, but would be redundant.

Random portfolios are useful when the constraints are different for the two strategies. For instance if one strategy is much less volatile than the other, then a comparison of returns is not especially appropriate. Each strategy can be mimicked by random portfolios, and the difference in daily p-values tested.

6 Constraint Evaluation

Another application of random portfolios is to get a sense of the usefulness of constraints that we put on the portfolio.

One of the constraints in the example was a maximum weight of 10%. A set of random portfolios were generated with the maximum weight constraint removed. Figure 15 shows the quantiles. Figure 16 compares the terminal wealth of the random portfolios with the 10% limit on the maximum weight with that of the random portfolios with no limit on the maximum weight. The two distributions are remarkably similar. Of particular interest is whether imposing the weight limit restricts the upper tail of the wealth. There is no evidence of that.

The weight constraint has a minimal effect on the random portfolios, so there remains the question of its effect on the strategy. The constraint avoids large losses from a single stock prediction being wrong, but also removes the possibility of a large gain from a single stock with a correctly large prediction. For the example alpha model, the constraint is undoubtedly useful since it not uncommonly gives a signal in the wrong direction. The appropriateness of this constraint appears to be largely a function of the quality of the alpha model.

Figure 15: Random portfolio quantiles (minimum, 5%, 10%, 25%, 50%, 75%, 90%, 95%, maximum) with no maximum weight constraint.

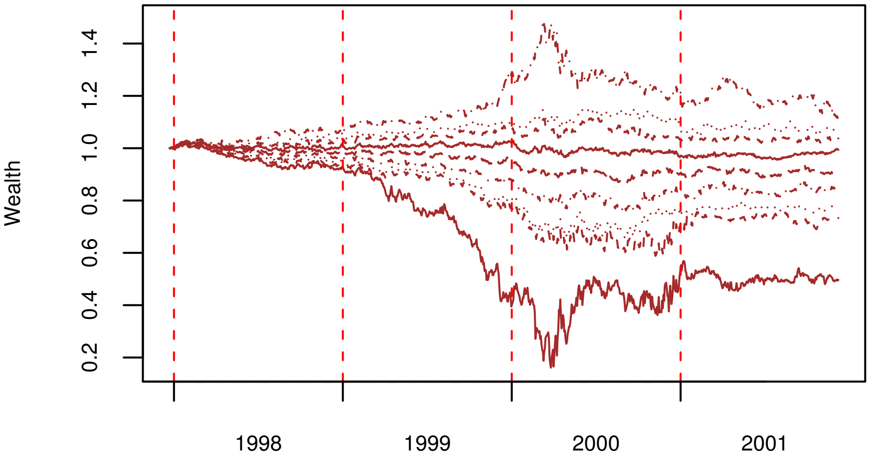
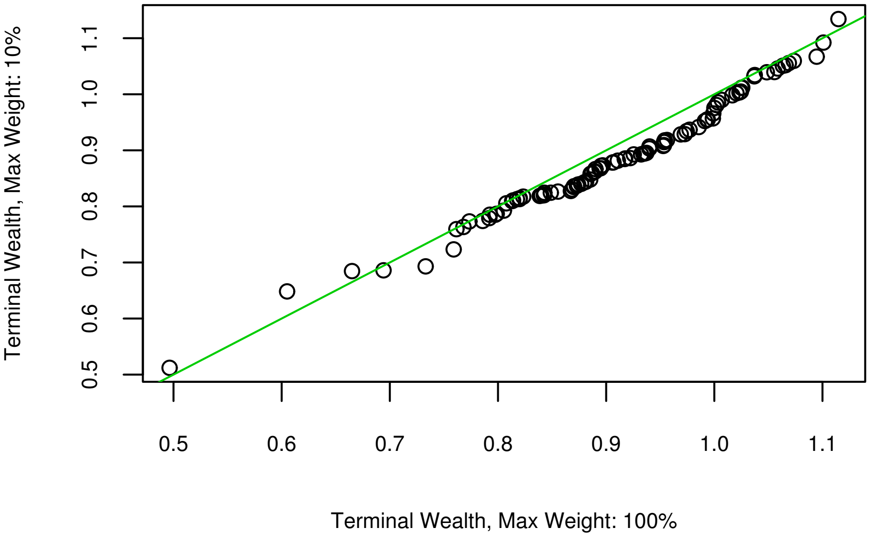


Figure 16: Comparison of the terminal wealth of the random portfolios with and without a maximum weight constraint.



7 Summary

There are three main problems when creating a trading strategy:

- Learning the strength of the prediction
- Evaluating the quality of the trading strategy
- Avoiding bias and false beliefs from data snooping

We've demonstrated a couple of statistical tests that deal with the first problem—the sign test and the Spearman correlation test.

Random portfolios can directly attack the second problem. They provide defensible and sensitive statements on the efficacy of a trading strategy. The results can be presented graphically with wealth curves or with p-value control charts.

Random portfolios also help some with the third problem. Random portfolios provide p-values, which can be adjusted to account for data snooping. A p-value of 0.001 is generally thought to be quite good. However, if you have tried a thousand different strategies and your best p-value is 0.001, then there is about a 63% probability of no value for the best strategy.

The more consistent your results across time and across different universes, the more confidence you can have that you are not just data snooping. It is standard practice to reserve a period of the most recent data to test the final strategy.

The focus here has been on researching a strategy before going live with it. However, p-value control charts—plots of p-values over time—are useful for live portfolios as well. The returns (or another measure of utility) over the recent past can be compared to those of a set of random portfolios. This gives instant feedback on the performance of the portfolio. A control chart can also be maintained that combines the results from a number of optimized portfolios with random starting points—this will show the current usefulness of the strategy itself.

References

[Burns, 2004] Burns, P. (2004). Performance measurement via random portfolios. Working paper, Burns Statistics, http://www.burns-stat.com/.

[Burns Statistics, 2005] Burns Statistics (2005). *POP Portfolio Construction User's Manual.* http://www.burns-stat.com.

[R Development Core Team, 2005] R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, http://www.r-project.org. ISBN 3-900051-07-0.